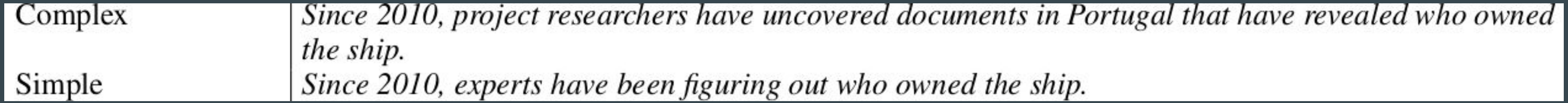
Controllable Text Simplification

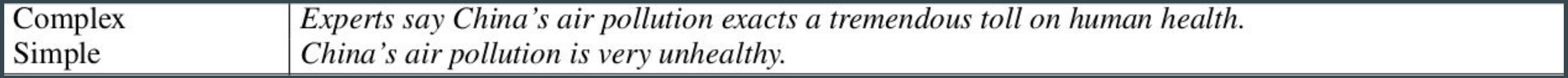
Adhiraj Deshmukh	(2021121012)
Aparna Agrawal	(2021121007)
Shreya Patil	(2021121009)

Problem Statement

- Text simplification refers to the process of **reducing the complexity** of a piece of text while **retaining its essential meaning**.
- The primary goal of text simplification is to make text more accessible and understandable to a broader audience, including individuals with limited literacy skills, non-native speakers, people with cognitive impairments, or those who are reading in challenging situations.
- It also helps with downstream natural language processing tasks, such as parsing, semantic role labelling, information extraction, and machine translation.

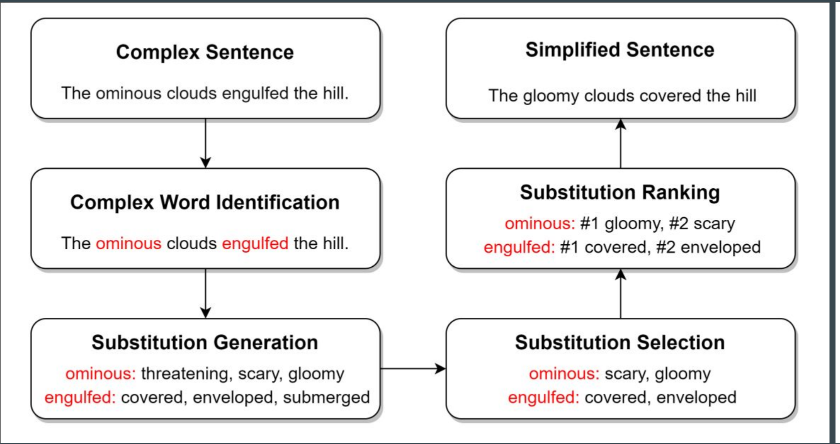


Complex	*Since 2010, project researchers have uncovered documents in Portugal that have revealed who owned the ship.*
Simple	*Since 2010, experts have been figuring out who owned the ship.*



Complex	*Experts say China's air pollution exacts a tremendous toll on human health.*
Simple	*China's air pollution is very unhealthy.*

Literature



Complex Sentence

The ominous clouds engulfed the hill.

Complex Word Identification

The ominous clouds engulfed the hill.

Substitution Generation

ominous: threatening, scary, gloomy
engulfed: covered, enveloped, submerged

Substitution Selection

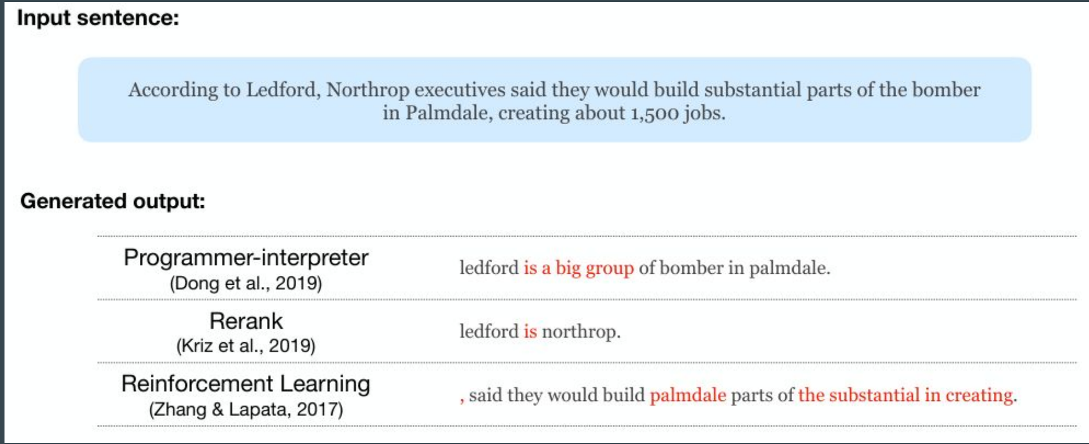
ominous: scary, gloomy
engulfed: covered, enveloped

Substitution Ranking

ominous: #1 gloomy, #2 scary
engulfed: #1 covered, #2 enveloped

Simplified Sentence

The gloomy clouds covered the hill



Input sentence:

According to Ledford, Northrop executives said they would build substantial parts of the bomber in Palmdale, creating about 1,500 jobs.

Generated output:

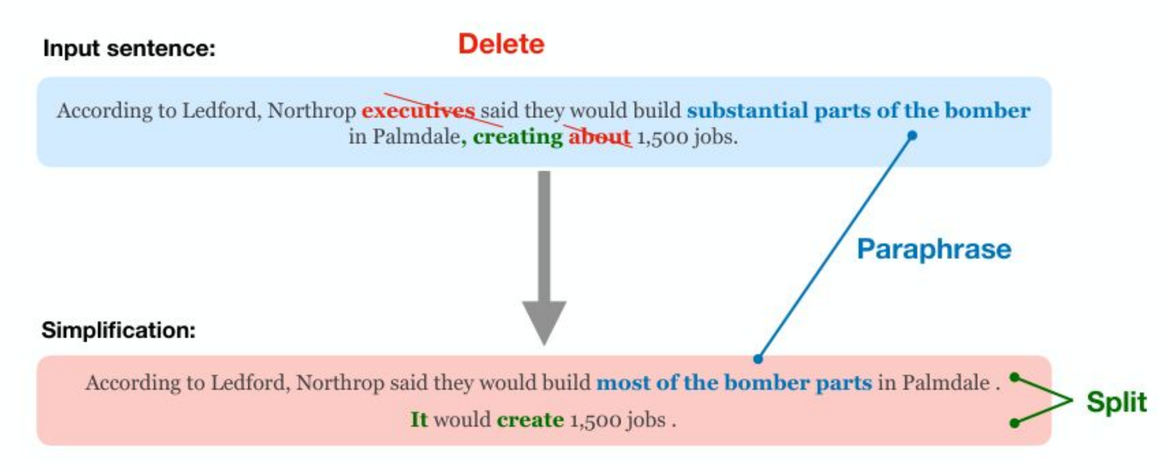
Programmer-interpreter (Dong et al., 2019)	ledford is a big group of bomber in palmdale.
Rerank (Kriz et al., 2019)	ledford is northrop.
Reinforcement Learning (Zhang & Lapata, 2017)	, said they would build palmdale parts of the substantial in creating.

- Most recent text simplification systems have adopted sequence-to-sequence models, *enhancing the fluency* of their output. However, these models primarily rely on **word deletion**, often resulting in very *short sentences* that **sacrifice the preservation of meaning**.

- While deleting words is a straightforward way to simplify sentences, it's not the most optimal or satisfying method. Human editors typically use a combination of techniques, including **deletion, paraphrasing, and breaking** sentences into smaller parts, to achieve better simplification.

- Another limitation of these end-to-end neural systems is the **lack of control**.

Literature



Input sentence:

According to Ledford, Northrop ~~executives~~ said they would build substantial parts of the bomber in Palmdale, creating ~~about~~ 1,500 jobs.

Simplification:

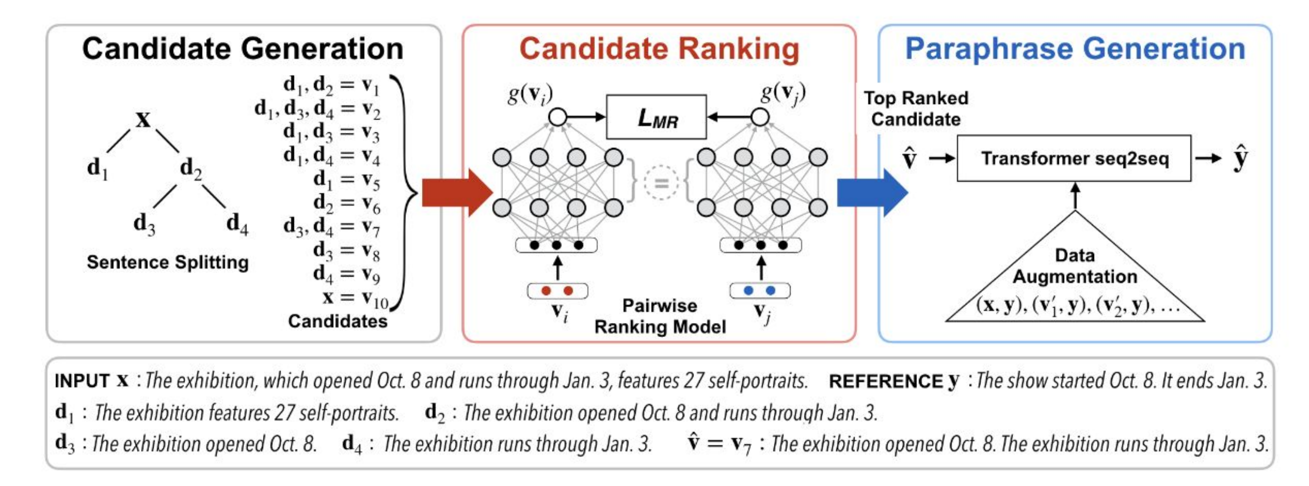
According to Ledford, Northrop said they would build most of the bomber parts in Palmdale .

It would create 1,500 jobs .

3 Primary Operations:

- **Delete**: Remove redundant or difficult words
- **Split**: Divide the text into multiple sentences to explicitly show the sequential nature of events or help decompose into simpler forms.
- **Paraphrase**: Replace the difficult or long phrases with simpler and shorter alternatives, respectively.

Paper’s Approach


Candidate Generation
$\mathbf{x}$
$\mathbf{d}_1$
$\mathbf{d}_2$
$\mathbf{d}_3$
$\mathbf{d}_4$
Sentence Splitting
$\mathbf{d}_1, \mathbf{d}_2 = \mathbf{v}_1$
$\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_4 = \mathbf{v}_2$
$\mathbf{d}_1, \mathbf{d}_3 = \mathbf{v}_3$
$\mathbf{d}_1, \mathbf{d}_4 = \mathbf{v}_4$
$\mathbf{d}_1 = \mathbf{v}_5$
$\mathbf{d}_2 = \mathbf{v}_6$
$\mathbf{d}_3, \mathbf{d}_4 = \mathbf{v}_7$
$\mathbf{d}_3 = \mathbf{v}_8$
$\mathbf{d}_4 = \mathbf{v}_9$
$\mathbf{x} = \mathbf{v}_{10}$
Candidates
Candidate Ranking
$g(\mathbf{v}_i)$
L_{MR}
$g(\mathbf{v}_j)$
=
$\mathbf{v}_i$
Pairwise Ranking Model
$\mathbf{v}_j$
Paraphrase Generation
Top Ranked Candidate
$\hat{\mathbf{v}}$
Transformer seq2seq
$\hat{\mathbf{y}}$
Data Augmentation
$(\mathbf{x}, \mathbf{y}), (\mathbf{v}'_1, \mathbf{y}), (\mathbf{v}'_2, \mathbf{y}), \ldots$

INPUT $\mathbf{x}$: *The exhibition, which opened Oct. 8 and runs through Jan. 3, features 27 self-portraits.* **REFERENCE $\mathbf{y}$** : *The show started Oct. 8. It ends Jan. 3.*

$\mathbf{d}_1$: *The exhibition features 27 self-portraits.* $\mathbf{d}_2$: *The exhibition opened Oct. 8 and runs through Jan. 3.*

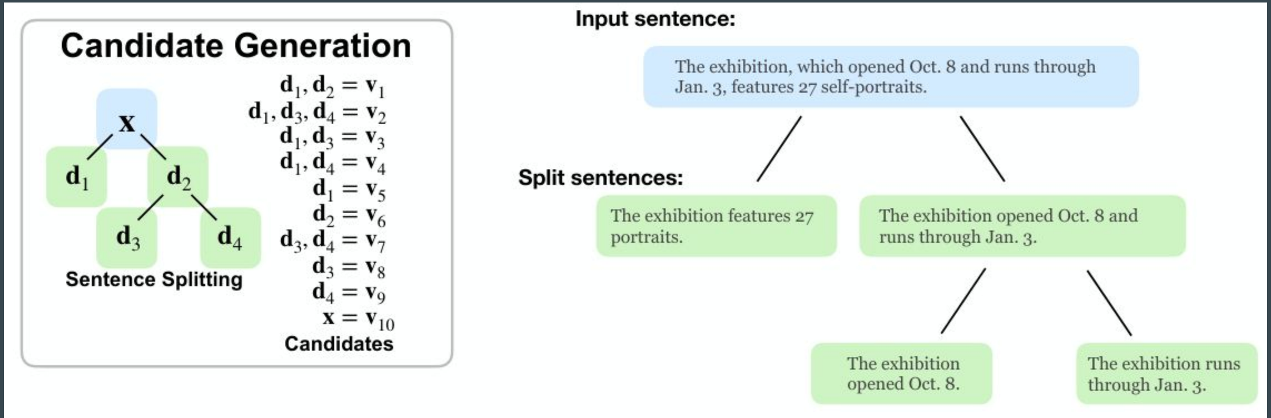
$\mathbf{d}_3$: *The exhibition opened Oct. 8.* $\mathbf{d}_4$: *The exhibition runs through Jan. 3.* $\hat{\mathbf{v}} = \mathbf{v}_7$: *The exhibition opened Oct. 8. The exhibition runs through Jan. 3.*

- Combines linguistically motivated syntactic rules with data-driven neural models to improve the diversity (**augmentation**) and **controllability**.

- **Decouples** the task to different components as they hypothesize that the seq2seq generation model will learn lexical and structural paraphrases more **efficiently**, when we offload some of the burden of sentence splitting, deletion and ranking decisions to separate components.

- Provides control over the 3 edit operations: **Deletion, Splitting and Paraphrasing**

Implementation: Candidate Generation



The exhibition, which opened Oct. 8 and runs through Jan. 3, features 27 self-portraits.

The exhibition features 27 portraits.

The exhibition opened Oct. 8 and runs through Jan. 3.

The exhibition opened Oct. 8.

The exhibition runs through Jan. 3.

- Uses an explicit linguistic rule based method, **DimSim** (Niklaus et al., 2019)
  - Best candidates are selected based on **compression ratio.** (avoiding extremely long and short)

- To increase the variety of candidates, a seq2seq Transformer model pre-trained on text simplification datasets is used to supplement DimSim.
  - To generate candidates, **Beam Search** is constrained to produce 10 outputs based on splitting, and 10 without splitting.
  - Then candidates that do not deviate too much from the source are selected based on **Jaccard Similarity**.

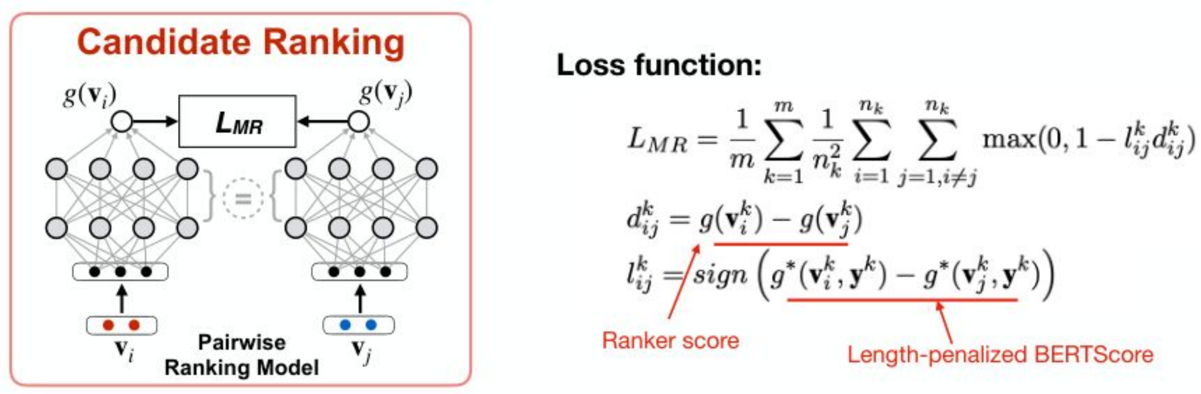
- We adopt a T5 model pre-trained on WikiAuto as our Neural Splitter and Deletion module.

Implementation: Candidate Generation

Challenges and Limitations:

- Explanation of constraints on splitting in beam search was left ambiguous, and thus we interpreted them as beam search on split outputs from DimSim module.
- The paper proposes the use of Neural Splitting module to supplement more candidate sentences, but there is no mechanism present while using this module to ensure generating strictly splitting and deletion based sentences.
- Thus, no guarantee on meaning preservation using **Jaccard Similarity**.

Implementation: Ranking



Loss function:

$$L_{MR} = \frac{1}{m}\sum_{k=1}^{m}\frac{1}{n_k^2}\sum_{i=1}^{n_k}\sum_{j=1,i\neq j}^{n_k}\max(0, 1 - l_{ij}^k d_{ij}^k)$$

$$d_{ij}^k = \underline{g(\mathbf{v}_i^k) - g(\mathbf{v}_j^k)}$$

$$l_{ij}^k = sign\left(\underline{g^*(\mathbf{v}_i^k, \mathbf{y}^k) - g^*(\mathbf{v}_j^k, \mathbf{y}^k)}\right)$$

Ranker score

Length-penalized BERTScore

- Neural Ranking model is used to score all the candidates that underwent splitting and deletion
  - The top ranked are feed to the paraphrasing module

- The scoring function is defined as **length-penalized BERTScore**
  - BERTScore is a metric of semantic similarity between 2 text
  - Penalty is present to favour candidates of similar lengths as reference

- The ranking model (MLP) is trained in a pairwise setup since BERTScore is relative rather than absolute similarity.
  - Features used for the ranking model are hand-crafted features (compression ratio, Jaccard similarity, etc.) that are vectorized using **Gaussian Binning**, and then concatenated together before feeding them to the model.

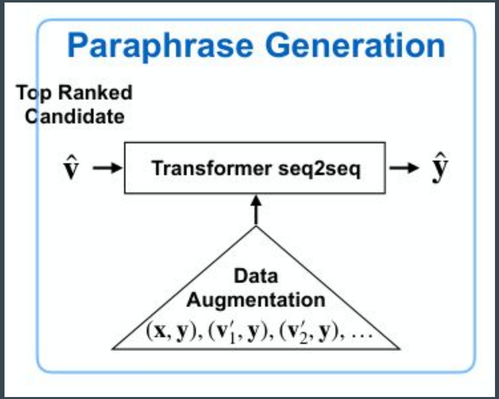
Implementation: Ranking

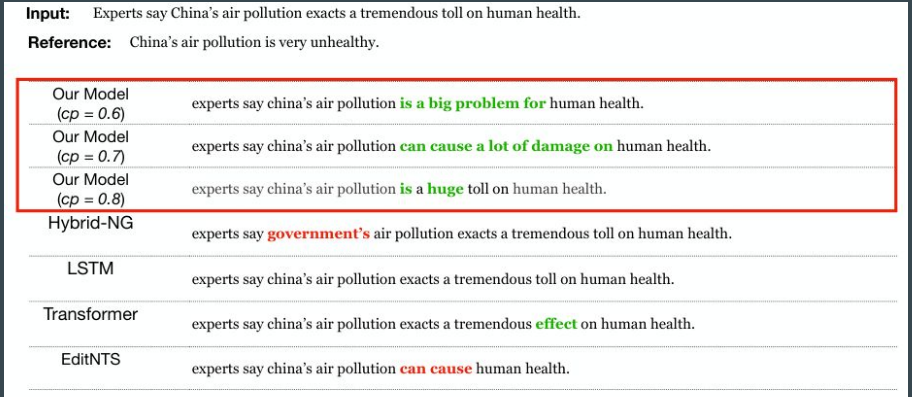
Challenges and Limitations:

- Features used in the ranking model that are based on rules used for Splitting and Deletion were not used due to the DimSim being **unable to extract** these rules.

- No clear motivation behind hand-crafted features have been provided and their impact on the ranking, which leaves a score for improvement in this part.
  - Thus, we tried experimenting with embeddings obtained from **SentenceBERT** and **Cosine Similarity** between the embeddings of source and candidate as features.
  - They perform slightly worse due to lack of information about the length of the sequence, and also take a lot more time to compute.

Implementation: Paraphrase Generation





Input: Experts say China's air pollution exacts a tremendous toll on human health.

Reference: China's air pollution is very unhealthy.

Our Model (*cp = 0.6*)	experts say china's air pollution **is a big problem for** human health.
Our Model (*cp = 0.7*)	experts say china's air pollution **can cause a lot of damage on** human health.
Our Model (*cp = 0.8*)	experts say china's air pollution **is** a **huge** toll on human health.
Hybrid-NG	experts say **government's** air pollution exacts a tremendous toll on human health.
LSTM	experts say china's air pollution exacts a tremendous toll on human health.
Transformer	experts say china's air pollution exacts a tremendous **effect** on human health.
EditNTS	experts say china's air pollution **can cause** human health.

- Paraphrase generation module is used to explicitly control the extent of lexical paraphrasing by specifying the percentage of words to be copied from the source sentence as a soft constraint.

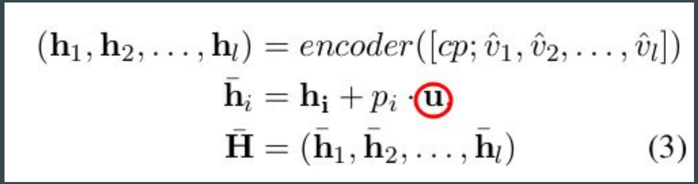
- The base model is just a BERT checkpoint pretrained on a text simplification dataset.

- Copy Control is incorporated by converting the parameter *cp* into a vector using **Gaussian binning** and appending it to the input sequence embeddings.
  - The transformer encoder then produces a sequence of hidden states, which are scaled by a learnt vector *U* according to the probability *pi* predicted the Copy Network.
  - Decoder uses this modified Hidden states to produce the final simplified sentence.

Implementation: Paraphrase Generation

Challenges and Limitations:

- We were unable to reliably train the model using the approach specified in the paper.
- Since the paper does not provide an intuitive explanation for some steps of the algorithm, we found it difficult to debug these parts of the code.
- As such we decided to adopt parts of an alternate approaches provided by Martin et al. (LREC 2020) and Cripwell et al., (Findings 2022)



$$\begin{aligned}(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_l) &= encoder([cp; \hat{v}_1, \hat{v}_2, \ldots, \hat{v}_l]) \\ \bar{\mathbf{h}}_i &= \mathbf{h_i} + p_i \cdot \boxed{\mathbf{u}} \\ \bar{\mathbf{H}} &= (\bar{\mathbf{h}}_1, \bar{\mathbf{h}}_2, \ldots, \bar{\mathbf{h}}_l) \qquad (3)\end{aligned}$$

Proposed Alternative - 1

- We take motivation from **Martin et al. (LREC 2020)**, which provides explicit control on simplification systems based on seq2seq models by using **discrete parametrization** mechanism.

- And we represent the Copy control parameter similarly by dividing the value space (0.0, 1.0] into multiple discrete buckets, which are represented by their corresponding special tokens, concatenated to the input sequence.

- While training, the token is selected using the ratio of number of common words between the input and reference sentences and total words, and then selecting the bucket corresponding to its value range.

- The intuition is very similar to the use of **prompt / prefix tuning**, where special tokens are learnt to provide **better representation** of the downstream task as compared to hard prompts.

Proposed Alternative - 2

- Another interesting approach that we encountered is **Cripwell et al., (Findings 2022)** which provides explicit control over simplification via **operation classification**.

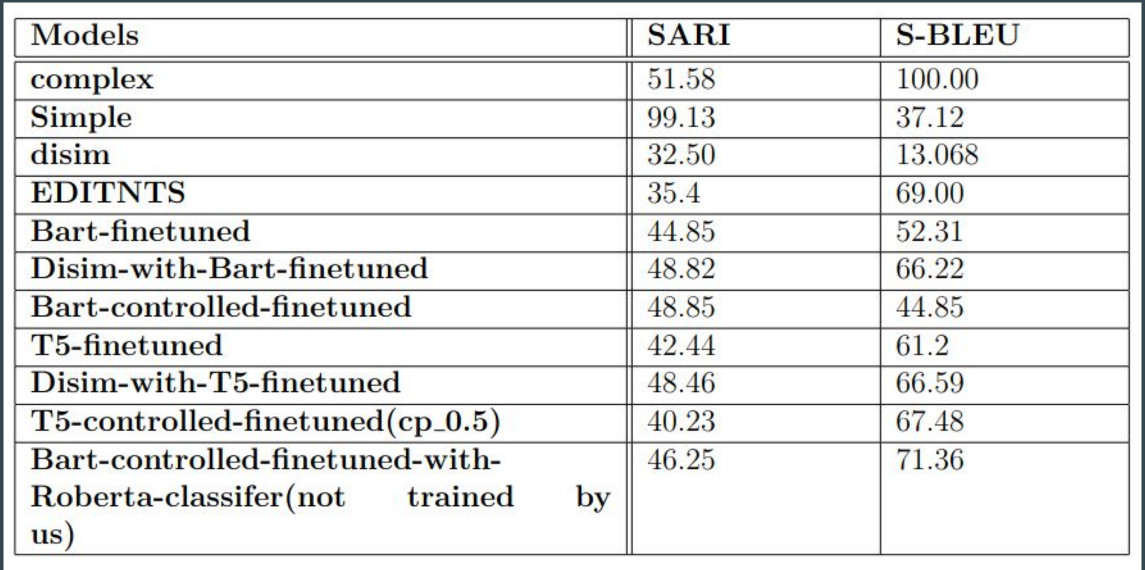
- They introduce a token which represents the class of the operation required for simplification of the input sentence

- This is similar to the previous approach, but we have **less fine-grained control** over the extent of the simplification caused by an operation of choice

- Thus this doesn’t serve as a direct alternative to the copy control module.

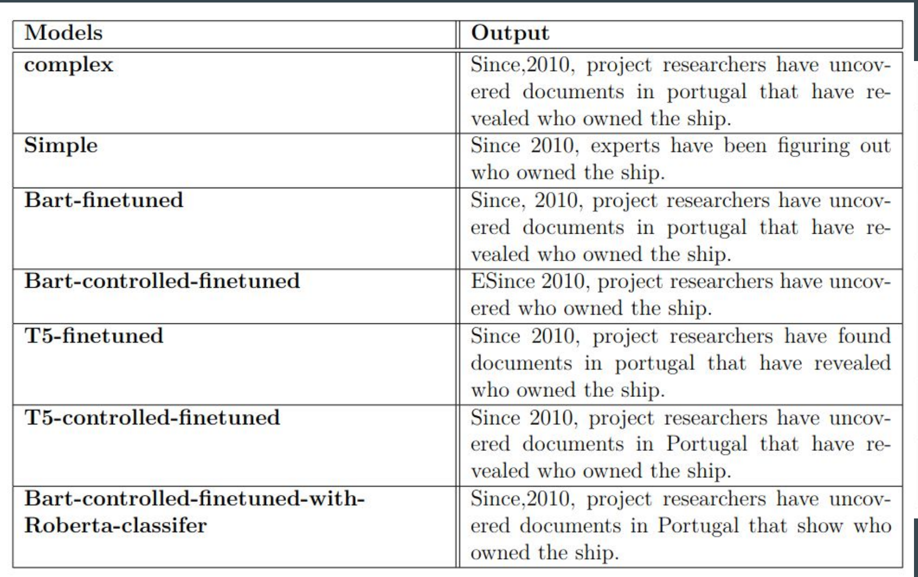
- This approach also requires curated operation **labels of the paired sentences** in the dataset for fine-tuning

Quantitative Results

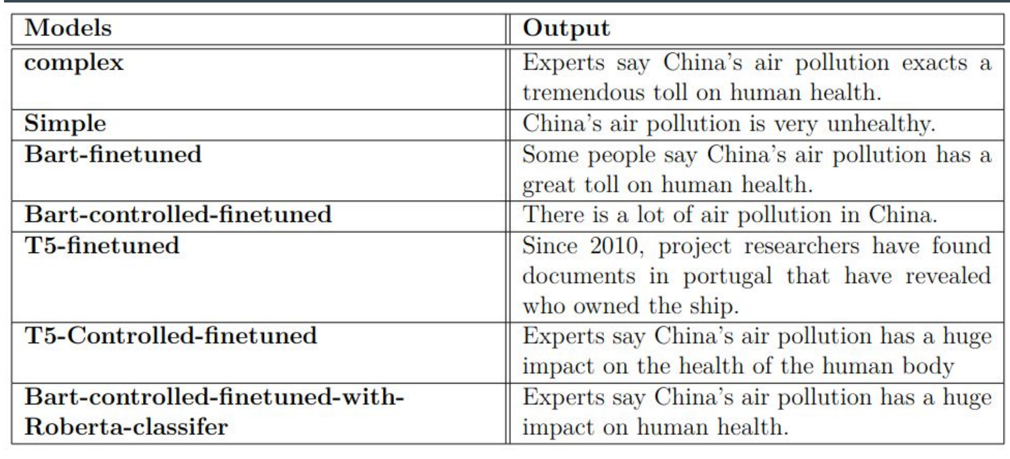


Models	SARI	S-BLEU
complex	51.58	100.00
Simple	99.13	37.12
disim	32.50	13.068
EDITNTS	35.4	69.00
Bart-finetuned	44.85	52.31
Disim-with-Bart-finetuned	48.82	66.22
Bart-controlled-finetuned	48.85	44.85
T5-finetuned	42.44	61.2
Disim-with-T5-finetuned	48.46	66.59
T5-controlled-finetuned(cp_0.5)	40.23	67.48
Bart-controlled-finetuned-with-Roberta-classifer(not trained by us)	46.25	71.36

Qualitative Results



Models	Output
complex	Since,2010, project researchers have uncovered documents in portugal that have revealed who owned the ship.
Simple	Since 2010, experts have been figuring out who owned the ship.
Bart-finetuned	Since, 2010, project researchers have uncovered documents in portugal that have revealed who owned the ship.
Bart-controlled-finetuned	ESince 2010, project researchers have uncovered who owned the ship.
T5-finetuned	Since 2010, project researchers have found documents in portugal that have revealed who owned the ship.
T5-controlled-finetuned	Since 2010, project researchers have uncovered documents in Portugal that have revealed who owned the ship.
Bart-controlled-finetuned-with-Roberta-classifer	Since,2010, project researchers have uncovered documents in Portugal that show who owned the ship.



Models	Output
complex	Experts say China's air pollution exacts a tremendous toll on human health.
Simple	China's air pollution is very unhealthy.
Bart-finetuned	Some people say China's air pollution has a great toll on human health.
Bart-controlled-finetuned	There is a lot of air pollution in China.
T5-finetuned	Since 2010, project researchers have found documents in portugal that have revealed who owned the ship.
T5-Controlled-finetuned	Experts say China's air pollution has a huge impact on the health of the human body
Bart-controlled-finetuned-with-Roberta-classifer	Experts say China's air pollution has a huge impact on human health.

More Results here: link

Any Questions ?